Representing discourse information for spoken dialogue generation

REPRESENTING DISCOURSE INFORMATION FOR SPOKEN DIALOGUE GENERATION*

Mark Steedman
Computer and Information Science, University of Pennsylvania
200 South 33rd Street
Philadelphia PA 19104-6389
(steedman@cis.upenn.edu)

ABSTRACT

Prosody and intonation convey important distinctions of "Information Structure", marking portions of the utterance as standing in relations to the surrounding discourse such as "theme" and "rheme", and marking relations of contrast between referring expressions and potential reference sets. The use of default intonation contours in standard "text-to-speech" applications can be quite successful, especially when the default pitch-accent assignments are moderated by "previous mention" information as an approximation to informational "given-ness" However, for some applications, particularly those involving dialogue rather than spoken monologue from written text, as well as ones involving systematic comparison among and coordination of alternatives, generation from richer meaning-representations can offer a more reliable alternative. Possible applications are inventory query, translating telephony for theme-marking languages like Japanese, personalised patient medical education, and animation of autonomous conversational agents.

1. INTRODUCTION

As a simple example, a query such as a, below, is most helpfully and intelligible answered with pitch-accents as indicated by capitals in b. Assignment of default final stress, as in c, sounds unnatural and will tend to reduce comprehension.

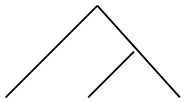
(1) a. Should I take the tablets with or without food?
b. You should take the tablets WITHOUT food.
c. #You should take the tablets without FOOD.

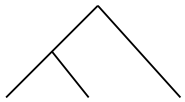
In this case, the strategem of destressing previously mentioned words will not help.

Steedman [9], Prevost and Steedman [8], and Prevost [7] discuss the alternative of representing the reference domain, including the theme or "open proposition" established by queries and other prior contexts, and including sets of alternative referents distinguished by various properties.

2. GRAMMAR AND INTONATION

On the basis of evidence from coordination and other constructions involving unbounded dependencies, a number of theories based on Categorial Grammar including Combinatory Categorial Grammar (CCG Steedman [9]) make the claim that substrings like *Anna married* are possible surface syntactic constituents. According to these theories, even such minimal sentences as *Anna married Manny* have two possible surface structures:

(2) a. 

b. 

More complex sentences like *Harry says that Anna married Manny* may have many surface structures for each reading.

Crucially, CCG provides a compositional semantics whereby both of the above analyses deliver the same logical form – say, $married'manny'anna'$ – the first by the application of a traditional predicate $\lambda y.married'manny'y$ with a subject $anna'$, and the other by the formation of a *non*-traditional predicate $\lambda x.married'x\ anna'$ and its application to an object $manny'$.

I shall not go into the details of CCG here, but the implications for the study of prosody should be obvious. One of the problems in defining the syntax-phonology interface with the simplicity which speech technological applications require is that intonation does not adhere to the tradition subject predicate division of the sentence. It too can bracket the sentence either way:

(3) Q: Well, what about MUSICALS? Who admires THEM?
A: (MARY) (admires MUSICALS).
H* L L+H* LH%

(4) Q: Well, what about MARY? What does SHE admire?
A: (MARY admires) (MUSICALS).
L+H* LH% H* LL%

The prosody is indicated formally using Pierrehumbert's notation, and informally with capitalisation and brackets. CCG offers a way to bring phrasal phonology and syntax closer together.

3. THEME AND RHEME

I have argued, following similar claims, implicit and explicit, by Jackendoff, Ladd, Gussenhoven , Selkirk, Rochemont, that the L+H* LH% contour in such examples marks the *theme* or topic (which in these contexts is the open proposition established by the question), while the H*LL% tune marks the rheme or comment (in these contexts, the answer). These objects correspond to λ-terms built by the CCG semantics. However, this proposal remains controversial, and Pierrehumbert and Hirschberg have related this tune

*Revised version of paper in the Proceedings of the International Symposium on Spoken Dialogue, held in conjunction with ICSLP-96, Philadelphia Oct. 1996. The research was supported by IRI91-17110 and IRI95-04372, ARPA grant no. N66001-94-C6043, and ARO grant no. DAAH04-94-G0426.

to a compositional semantics for intonational tunes that is based on scalar values on dimensions such as certainty concerning relevance and degree of commitment to belief revision [6, p.294-297]. According to their account, the L+H* pitch accent is used "to convey that the accented item – and not some alternative related item – should be mutually believed" (p.296).

The following minimal pair of dialogues will be helpful in deciding between these alternatives, because it appears at first glance to raise problems for both.

(5) Q: Does Mary love opera?
A: Mary admires MUSICALS.
H* LL%

(6) Q: Does Mary love opera?
A: Mary admires MUSICALS.
L+H* LH%

In the first example, the entire response is marked with the H*LL% tune that we have identified as marking the rheme, constituting what the speaker believes the hearer needs. Depending on the context, the speaker may thereby be committed by the usual Gricean principles to a number of conversational implicatures. For example, if admiring musicals entails hating opera, then this response implicates denial. If on the other hand admiring musicals entails loving opera, then affirmation is implicated. Either way, the speaker's intonation only commits them to the claim that the information given is adequate for the hearer's needs, rather than to a particular belief.

The second example appears at first glance to be almost equivalent. In particular, the possibilities for conversational implicature of either affirmation or denial seem identical, so it does not seem as though the alternatives are uniquely evoked by the L+H* pitch accent. Both pitch accents are constrastive – in fact, all pitch accents are, as Pierrehumbert and Hirschberg point out ([6, p.288-9]) – and this is where the alternatives are evoked.[1]

Any difference seems to lie in the degree of commitment concerning relevance to the question at issue that the speaker brings to the utterance of the information. Since in other respects the two utterances seem similar, there is a temptation to believe that the L+H* LH% tune in this case might mark a rheme, rather than a theme. However, it is also possible that what the respondent has actually done is to offer the proposition as an alternative theme, leaving the other party to supply a rheme. If so, the effect of not taking responsibility for a rheme in this utterance is likely to be that of to conversationally implicating a lack of confidence in either the relevance of the theme or the certainty of the inference that might be drawn. But that would not be a matter of literal or conventional meaning of the utterance itself.

This is essentially the analysis proposed by Ladd [5, p152-6] who relates all uses of "fall-rise" contours including this one to the function of establishing a set of alternatives established by the preciding speaker's question – a notion which we have identified with the notion of theme, and which seems likely to translate rather naturally into the "alternatives semantics" of Karttunen and Peters and Rooth.

[1] These examples also demonstrate that the accented material and the alternatives that it evokes need not have been mentioned or evoked by the first speaker. The respondent may cause previously unmentioned and unevoked sets to be "accomodated" by the other, in Lewis' sense of the term.

Further support for the claim that the L+H* LH% tune marks theme, and that the effect of lack of committment arises by conversational implicature can be found in the fact that this intonation remains appropriate when the step of inference that generates the rheme itself is explicitly spelled out, as in the following deliberately exaggerated example, in which *admiring musicals* is necessarily distinct from the rheme:

(7) Q: Does Mary love opera?
A: Well, she admires MUSICALS,
L+H* LH%
And people who admire MUSICALS always love OPERA.
L+H* LH% H*LL%
So I am sure she loves OPERA.
H*LL%

(Note that *admiring musicals* in the first conjunct could equally well be uttered with an H*L% rheme accent, but in the second it really has to be marked as a theme. Under most circumstances the first and third conjuncts could be omitted entirely, as being implicated by the second.)

Example 6 is closely related to examples discussed by Ward and Hirschberg, who notate a similar tune with an L*+H pitch accent – cf. Pierrehumbert and Hirschberg [6, p.295, ex. 26].

(8) A: Harry's such a klutz.
B: He's a good BADMINTON player.
L*+H LH%

They describe the L*+H LH% tune (which Ladd subsumes under fall-rise, and does not distinguish from L+H* LH%) in terms that are remarkably similar to those used above to describe the L+H* LH% tune in example 6, although they do not invoke the theme/rheme distinction. Thus they describe the above utterance as expressing "uncertainty about whether being a good badminton player provides relevant information about degrees of clumsiness."

This is very close to the claim of the present theory, but the intonation in Ward and Hirschberg's example, 8 above, is consistent with an elaboration along the same lines as 7, supplying a rheme to the effect that badminton players are invariably dexterous, so again it seems unlikely that this uncertainty is a matter of literal or conventional meaning. (It also makes it unclear that it is *degrees* of clumsiness that are at issue, rather than clumsiness *tout court*.)

The L+H* and L*+H pitch accents are hard to tell apart, both subjectively and instrumentally, although Pierrehumbert & Steele present experimental support for the claim that the distinction is categorical. Pierrehumbert and Hirschberg note that their discourse functions are closely related. Perhaps all of the L+H pitch accents discussed up to this point should be notated as L*+H. Or perhaps the distinction between the various fall-rise tunes is allophonic as Liberman and Ladd would have it, and the supposed distinction in function between L+H* and L*+H is in fact on a continuum related to degree of contrast associated with the pitch accent. We will remain agnostic on the question of distinctions within the fall-rise family of tunes, using the L+H* LH notation for all types, without prejudice to these authors' claim.

It is only appropriate to mark the theme with an L+H* pitch accent when it stands in contrast to an preceding different theme. If the topic to which a theme refers is unambiguously established, it is

common to find that the theme is deaccented throughout and (if it is utterance-initial) without any boundary, as in the following:

(9) Q: What does Mary admire?
A: (Mary admires) (MUSICALS).
H* LL%

We would be missing an important semantic generalisation if we failed to note that examples 4 and 9 are identical in information structure as far as the theme-rheme division goes. We shall therefore need to distinguish the "marked" theme in the former from the "unmarked" theme in the latter. Unmarked intonation is always ambiguous as to information structure. In the following context, the same contour will have the information structure of 3:

(10) Q: What is Mary like?
A: (Mary) (admires MUSICALS).
H* LL%

4. FOCUS AND GROUND

The possibility of such unmarked themes, lacking any pitch accent, draws attention to a second independent dimension to discourse information structure that affects intonational tune. In example 4, the L+H* LH% tune is spread across the entire substring of the sentence corresponding to the theme in the above sense – that is, over the substring *Mary admires*.[2] In 3, the same tune L+H* LH% is confined to the object of the theme *admires musicals*, because the intonation of the original question indicates that admiring musicals *as opposed to admiring something else* is the new topic or theme. In 9 and 10, there is no L+H* LH% tune at all.

The position of the pitch accent in the phrase has to do with a further dimension of information structure *within both theme and rheme*, corresponding to a distinction between *the interesting part(s)* of either information unit, and the rest. Halliday, who was probably the first to identify the orthogonal nature of these two dimensions, called it "new" information, in contrast to "given" information. The term "new" is not entirely helpful, since (as Halliday was aware), the relevant part of the theme need not be novel to the discourse, as in the examples to hand. We will follow the phonological literature and Prevost [7] in calling the information marked by the pitch accent the "focus", distinguishing theme-focus and rheme-focus where necessary, and use the term "ground" for the part unmarked by pitch-accent or boundary. Again there are other taxonomies, with most of which the present proposal is compatible.[3]

The following example serves to illustrate the full range of possibilities for the distribution of focus and ground within theme and rheme.

(11) Q: I know that Mary envies the woman who directed the musical.
But who does she ADMIRE?
A: (Mary ADMIRES) (the man who WROTE the musical)

[2] An equally felicitous prosody in which the theme tune is confined to *Mary* is dicussed in the 1991 paper.

[3] It is important to know that the term "focus" is used in the literature in several other conflicting ways.

Here the theme is *Mary admires*, where only *admires* is emphasised because the previous theme was also about Mary. The rheme is *the man who wrote the musical*, where only *wrote* is contrasted.

5. DYNAMICS OF CONTEXT

The theme has the character of a referring expression, much like the interpretation of a definite NP. When a *wh*- question is used to establish a new theme-referent, the speaker typically assigns the rheme-tune H*LL% to the residue of *wh*-movement, as in the example 4, repeated here with intonation indicated.

(12) What does MARY admire?
H* LL%

The use of a rheme to set up a referent to which the theme of a subsequent utterance may refer is quite general. The rheme in the response in 4 has the same effect:

(13) MARY admires MUSICALS
L+H* LH% H* LL%

The rheme *musicals* can establish a thematic referent to which the theme of a subsequent utterance refers, as in the following example:

(14) MUSICALS are MARVELOUS!
L+H* LH% H* LL%

In the light of these observations, it will be helpful to think of the semantics and pragmatics of information structure in terms of a simple discourse model that could conveniently be more or less directly realised in the programming language Prolog. The context consists of a small set of facts, represented by terms, such as $admire'musicals'mary'$. Queries of the form $admire'musicals'mary'$, and $admire'corduroy'mary'$ either succeed or fail with respect to this database. λ-terms can be thought of as implicitly existentially quantified queries that succeed by constructively binding x to *musicals*. The database should be thought of as including such λ-terms defining one or more potential thematic referents of the discourse, to which the theme of subsequent utterances refer. A model of this kind has been investigated in some detail and implemented computationally by Prevost [7].

In such a framework, we can think of the theme of an utterance as *updating* or side-effecting the context or discourse model. It can be characterised, following Jacobs and Krifka, as causing one or more existing referents or facts of the form $(theme'\lambda\ldots)$ to be *retracted* or removed from the context model, and causing a new thematic referent or fact to be *asserted* or added. If the theme is unmarked by any accent, then it will simply be the corresponding thematic referent that is retracted and asserted. Unless a fact of the appropriate form is already present in (or is at least consistent with) the context, the first of these effects will precipitate a failure of the discourse. Otherwise, the thematic referent will be reasserted.

The above examples show that the rheme should also update the context with a new thematic referent. However, it does not cause any existing thematic referents to be retracted (although we shall see that it may have other effects on the database, via the entailments and implicatures discussed above.)

The isolated rhemes and themes discussed in connection with examples 5 and 6 then work as follows. In both cases, the rheme of the yes-no question adds a theme $theme'(love'mary'opera')$ to the facts

making up the the recipient's context. They then construct the corresponding query, and evaluate it with respect to the context. If the query immediately succeeds, or fails altogether, then it is appropriate to respond with a direct yes or no. If the query succeeds but a step of inference involving the respondent's rule that *everyone who admires musicals loves opera*, and the respondent's knowledge that *Mary admires musicals*, is needed to establish the answer, then one of the following cases may apply. If the respondent has reason to believe that the questioner knows neither the rule, nor the truth of the premise, then the respondent should state them both, as in the extended example 7. On the other hand, if the respondent has reason to believe that the recipient knows the rule, but not the premise, then they should respond either as in 5 or as in 6. If they have reason to believe that this is the only relevant difference between the questioners knowledge and their own, then stating the premise as a rheme, as in 5, is appropriate, since they can sincerely claim that it it is everything the questioner needs. But if they have reason to suspect that there may be other differences, and therefore cannot sincerely commit to this inference going through for the questioner, then they should mark the premise as a theme, as in 6, and leave the questioner to decide whether they can get a rheme out of it. Such reasoning about knowledge and belief can be expressed in modal logics of various kinds. While inference with such logics is in general intractable, Matthew Stone [10], building on work by Wallen, has identified a tractable proof algorithm for the subclass representing D4, S4, or T necessity in the absence of possibility and classical negation. These logical fragments are applicable to the kind of reasoning sketched above, and were motivated by work on generation of spoken conversation between autonomous animated humanoid agents (Cassell et al. [2]).

The exact form of the retracted and/or asserted informational referents in all of the above examples is dependent upon the location of focus and pitch accents in the utterance, and is determined in a manner discussed by Prevost, who gives an algorithm for assigning pitch accents to lexical items, building on work by Dale and Haddock.

6. MEANING-TO-SPEECH APPLICATIONS

Non-semantic text-to-speech with default intonation remains the technique of choice for domain-independent speech generation for tasks where errors can be tolerated, especially when augmented with recent mention heuristics (Terken and Hirschberg 1994, Hirschberg 1990), and it provides the baseline against which more costly alternatives are evaluated. However, for tasks in which the domain and the alternatives sets can be modelled, especially those in which errors cannot be tolerated, the techniques that we are investigating offer an alternative. One task that is currently being investigated by Komagata at Penn is spoken translation from languages which express themes morphologically, as Japanese does with *-wa*, although the mapping to English intonation is not direct – cf. Kuno [4].

Another application that we are currently investigating in collaboration with Johanna Moore's group at the University of Pittsburgh is the provision of interactive spoken patient-specific medical information. This domain is characterised by a large number of contrastive sets of medical conditions, symptoms, treatments, and so on. The domain can be modelled using standard knowledge representations, and can be used to generate informative text. For example, Consider the following two paragraphs taken from the output of the Pittsburgh MIGRAINE system (Carenini, Mittal, and Moore [1]):

(15) Drugs for prophylactic treatment are intended to reduce or prevent further migraine attacks. These are drugs that you must take every day, whether or not you have a migraine. In general, prophylactic treatment is suitable for patients with frequent migraines. The most common side effects of Nortriptyline are dry mouth and drowsiness.
Drugs for analgesic treatment are intended to reduce or relieve the pain of migraine headaches. These are drugs that you must take when headache occurs. In general, analgesic treatment is suitable for patients with infrequent or mild headaches. Some patients experience side-effects with Ibuprofen such as stomach pain and discomfort.

A patient in interaction with this system by voice or mouse might encounter these two paragraphs in either order. But whichever paragraph is uttered second by the system, it should bear a very different intonation from the first if the passage is not to be wearisome or hard to follow. (For example, phrases like "These are drugs . . . " should get a series of downstepped H* on nouns like *drugs* etc. the first time, and L+H* on *These* and stress on only the last noun the second time.) While these paragraphs were in fact written by a human, the domain is sufficiently constrained as to fall within the scope of the techniques investigated by Prevost (this conference). As the diagnostic inferential power behind such systems increases, so will the benefits of generating speech directly from the underlying knowledge, rather than by default or by canning a version for each context.

7. REFERENCES

1. Carenini, G. V. Mittal & J. Moore 1994. Generating patient-specific interactive natural language explanations, *Proceedings of SCAMC 94*.
2. Cassell, J., C. Pelachaud, N. Badler, M Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, & M. Stone), 1994. Animated conversation, *Proceedings of the ACM SIGRAPH '94 Conference*, Orlando FL, 413-420.
3. Hirschberg, J. 1993, Pitch accent in context. *Artificial Intelligence*, 63, 305-340.
4. Kuno, Susumo. 1973. *The Structure of the Japanese Language*, MIT Press, Cambridge MA.
5. Ladd, D.R. 1980. *The Structure of Intonational Meaning*, Indiana University Press, Bloomington IN.
6. Pierrehumbert, J. & J. Hirschberg, 1990. The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, & M. Pollack (eds.), *Intentions in Communication*, MIT Press Cambridge MA, 271-312.
7. Prevost, S. 1995. *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation*. Ph.D. thesis, University of Pennsylvania, IRCS report no. 96-01.
8. Prevost, S. & M. Steedman. 1994. Specifying intonation from context for speech synthesis, *Speech Communication*, 15, 139-153.
9. Steedman, M. 1991, Structure and intonation, *Language*, 67, 262-296.
10. Stone, M. 1996. Efficient tree construction for reasoning about necessity, ms. U. Penn.
11. Terken and Hirschberg, 1994. Deaccentuation of words representing 'given' information, *Language and Speech*, 37, 125-145.